

ROBUST RANK CORRELATION BASED SCREENING

BY GAORONG LI¹, HENG PENG², JUN ZHANG³ AND LIXING ZHU⁴

*Beijing University of Technology, Hong Kong Baptist University,
 Shenzhen University and Hong Kong Baptist University*

Independence screening is a variable selection method that uses a ranking criterion to select significant variables, particularly for statistical models with nonpolynomial dimensionality or “large p , small n ” paradigms when p can be as large as an exponential of the sample size n . In this paper we propose a robust rank correlation screening (RRCS) method to deal with ultra-high dimensional data. The new procedure is based on the Kendall τ correlation coefficient between response and predictor variables rather than the Pearson correlation of existing methods. The new method has four desirable features compared with existing independence screening methods. First, the sure independence screening property can hold only under the existence of a second order moment of predictor variables, rather than exponential tails or alikeness, even when the number of predictor variables grows as fast as exponentially of the sample size. Second, it can be used to deal with semiparametric models such as transformation regression models and single-index models under monotonic constraint to the link function without involving nonparametric estimation even when there are nonparametric functions in the models. Third, the procedure can be largely used against outliers and influence points in

Received March 2012; revised June 2012.

¹Supported in part by the NNSF (11101014) of China, the Specialized Research Fund for the Doctoral Program of Higher Education of China (20101103120016), PHR (IHLB, PHR20110822), the Training Programme Foundation for the Beijing Municipal Excellent Talents (2010D005015000002) and the Fundamental Research Foundation of Beijing University of Technology (X4006013201101).

²Supported in part by CERG grants from the Hong Kong Research Grants Council (HKBU 201610, HKBU 201809 and HKBU 202012), FRG grants from Hong Kong Baptist University (FRG/10-11/024 and FRG/11-12/130) and a grant from National Nature Science Foundation of China (NNSF 11271094).

³Supported in part by the NNSF (11101157) of China.

⁴Supported in part by a grant from the Research Grants Council of Hong Kong, and an FRG grant from Hong Kong Baptist University.

AMS 2000 subject classifications. Primary 62J02, 62J12; secondary 62F07, 62F35.

Key words and phrases. Variable selection, rank correlation screening, dimensionality reduction, semiparametric models, large p small n , SIS.

This is an electronic reprint of the original article published by the Institute of Mathematical Statistics in *The Annals of Statistics*, 2012, Vol. 40, No. 3, 1846–1877. This reprint differs from the original in pagination and typographic detail.

the observations. Last, the use of indicator functions in rank correlation screening greatly simplifies the theoretical derivation due to the boundedness of the resulting statistics, compared with previous studies on variable screening. Simulations are carried out for comparisons with existing methods and a real data example is analyzed.

1. Introduction. With the development of scientific techniques, ultra-high dimensional data sets have appeared in diverse areas of the sciences, engineering and humanities; Donoho (2000) and Fan and Li (2006) have provided comprehensive reviews. To handle statistical problems related to high dimensional data, variable/model selection plays an important role in establishing working models that include significant variables and exclude as many insignificant variables as possible. A very important and popular methodology is shrinkage estimation with penalization, with examples given of bridge regression [Frank and Friedman (1993), Huang, Horowitz and Ma (2008)], LASSO [Tibshirani (1996), van de Geer (2008)], elastic-net [Zou and Hastie (2005)], adaptive LASSO [Zou (2006)], SCAD [Fan and Li (2001), Fan and Peng (2004), Fan and Lv (2011)] and Dantzig selector [Candes and Tao (2007)]. When irrerepresentable conditions are assumed, we can guarantee selection consistency for LASSO and Dantzig selector even for “large p , small n ” paradigms with nonpolynomial dimensionality (NP-dimensionality). However, directly applying LASSO or Dantzig selector to ultra-high dimensional modeling is not a good choice because the irrerepresentable conditions can be rather stringent in high dimensions; see, for example, Lv and Fan (2009) and Fan and Lv (2010).

Fan and Lv (2008) proposed another promising approach called sure independence screening (SIS). This methodology has been developed in the literature by researchers recently. Fan and Song (2010) extended SIS to ultra-high dimensional generalized linear models, and Fan, Feng and Song (2011) studied it for ultra-high dimensional additive models. Moreover, based on the idea of dimension reduction, Zhu et al. (2011) suggested a model-free feature screening method for most generalized parametric or semiparametric models. To sufficiently use the correlation information among the predictor variables, Wang (2012) proposed a factor profile sure screening method for the ultra-high dimensional linear regression model. Different from existing methods with penalization, SIS does not use penalties to shrink estimation, but ranks the importance of predictors by correlations between response and predictors marginally for variable/model selection. To perform the ranking, Pearson correlation is adopted; see Fan and Lv (2008). For NP-dimensionality, the tails of predictors need to be nonpolynomially light. This is also the case for other shrinkage estimation methods such as the LASSO and Dantzig selector. Moreover, to use more information among the predictor variables to make a sure screening such as Wang (2012), or to apply

the sure screening method to more general statistical models such as Zhu et al. (2011), more restrictive conditions, such as the normality assumption [Wang (2012)] or the linearity and moment conditions [Zhu et al. (2011)], need be imposed on the predictor variables. To further improve estimation efficiency, Fan and Lv (2008) suggested a two-stage procedure. First, SIS is used as a fast but crude method of reducing the ultra-high dimensionality to a relatively large scale that is smaller than or equal to the sample size n ; then, a more sophisticated technique can be applied to perform the final variable selection and parameter estimation simultaneously. Note that for linear models, the SIS procedure also depends on the explicit relationship between the Pearson correlation and the least squares estimator [Fan and Lv (2008)]. For generalized linear models, Fan, Samworth and Wu (2009) and Fan and Song (2010) selected significant predictors by sorting the corresponding marginal likelihood estimator or marginal likelihood. That method can be viewed as a likelihood ratio screening, as it builds on the increments of the log-likelihood. The rate of p also depends on the tails of predictors. The lighter the tails are, the faster the rate of p can be. Xu and Zhu (2010) also showed for longitudinal data that when only the moment condition is assumed, the rate of p cannot exponentially diverge to infinity unless moments of all orders exist.

For other semiparametric models such as transformation models and single-index models, existing SIS procedures may involve nonparametric plug-in estimation for the unknown transformation or link function. This plug-in may deteriorate the estimation/selection efficiency for NP-dimensionality problems. Although the innovative sure screening method proposed by Zhu et al. (2011) can be applied to more general parametric or semiparametric models, as commented above, the much more restrictive conditions are required for the predictor variables. Zhu et al. (2011) imposed some requirements for the tail of the predictor variables which further satisfy the so-called linearity condition. This condition is only slightly weaker than elliptical symmetry of the distribution of the predictor vector [Li (1991)]. It is obvious that their sure screening method does not have the robust properties as the proposed method in this paper has. Further, when the categorical variables do involve the ultra-high dimensional predictor vector, the restrictive conditions on the predictor variables hinder the model-free feature screening method to apply directly. On the other hand, such a model-free feature screening method is based on slice inverse regression [SIR, Li (1991)]. It is well known that SIR is not workable to the model with symmetric regression function; see Cook and Weisberg (1991).

We note that the idea of SIS is based on Pearson correlation learning. However, the Pearson correlation is not robust against heavy tailed distributions, outliers or influence points, and the nonlinear relationship between response and predictors cannot be discovered by the Pearson correlation. As

suggested by Hall and Miller (2009) and Huang, Horowitz and Ma (2008), independence screening could be conducted with other criteria. For correlation relationships, there are several measurements in the literature, and the Kendall τ [Kendall (1938)] is a very commonly used one that is a correlation coefficient in a nonparametric sense. Similar to the Pearson correlation, the Kendall τ also has wide applications in statistics. Kendall (1962) gave an overview of its applications in statistics and showed its advantages over the Pearson correlation. First, it is robust against heavy tailed distributions: see Sen (1968) for parameter estimation in the linear regression model. Second, the Kendall τ is invariant under monotonic transformation. This property allows us to discover the nonlinear relationship between the response and predictors. For example, Han (1987) suggested a maximum rank correlation estimator (MRC) for the transformation regression model with an unknown transformation link function. Third, the Kendall τ based estimation is a U-statistic with a bounded kernel function, which provides us a chance to obtain sure screening properties with only a moment condition. Another rank correlation is the Spearman correlation [see, e.g., Wackerly, Mendenhall and Scheaffer (2002)]. The Spearman rank correlation coefficient is equivalent to the traditional linear correlation coefficient computed on ranks of items [Wackerly, Mendenhall and Scheaffer (2002)]. The Kendall τ distance between two ranked lists is proportional to the number of pairwise adjacent swaps needed to convert one ranking into the other. The Spearman rank correlation coefficient is the projection of the Kendall τ rank correlation to linear rank statistics. The Kendall τ has become a standard statistic with which to compare the correlation between two ranked lists. When various methods are proposed to rank items, the Kendall τ is often used to measure which method is better relative to a “gold standard.” The higher the correlation between the output ranking of a method and the “gold standard,” the better the method is. Thus, we focus on the Kendall τ only. More interestingly, the Kendall τ also has a close relationship with the Pearson correlation, particularly when the underlying distribution of two variables is a bivariate normal distribution (we will give the details in the next section). As such, we can expect that a Kendall τ based screening method will benefit from the above mentioned advantages to be more robust than the SIS.

The reminder of this paper is organized as follows. In Section 2 we give the details of the robust rank correlation screening method (RRCS) and present its extension to ultra-high dimensional transformation regression models. In Section 3 the screening properties of the RRCS are studied theoretically for linear regression models and transformation regression models. In Section 4 an iterative RRCS procedure is presented. We also discuss RRCSs application to generalized linear models with NP-dimensionality. Numerical studies are reported in Section 5 with a comparison with the SIS. Section 6 concludes the paper. A real example and the proofs of the main results can be found in the supplementary material for the paper [Li et al. (2012)].

2. Robust rank correlation screening (RRCS).

2.1. *Kendall τ and its relationship with the Pearson correlation.* Consider the random vectors $(X_i, Y_i), i = 1, 2, \dots, n$, and the Kendall τ rank correlation between X_i and Y_i is defined as

$$(2.1) \quad \tau = \frac{1}{n(n-1)} \sum_{i \neq j}^n \text{sgn}(X_i - X_j) \text{sgn}(Y_i - Y_j).$$

Given this definition, it is easy to know that $|\tau|$ is invariant against the monotonic transformation of X_i or Y_i . Furthermore, if (X_i, Y_i) follows a bivariate normal distribution with mean zero and the Pearson correlation ρ , it can be shown that [Huber and Ronchetti (2009)]

$$E(\tau) = \frac{2}{\pi} \arcsin \rho.$$

In other words, when (X_i, Y_i) follows bivariate normal distribution, the Pearson correlation and Kendall τ have a monotonic relationship in the following sense. If $|\rho| > c_1$ for a given positive constant c_1 , then there exists a positive constant c_2 such that $|E(\tau)| > c_2$, and if and only if $\rho = 0$, $E(\tau) = 0$. Such a relationship helps us to obtain the sure independence screening property for linear regression models under the assumption of Fan and Lv (2008) without any difficulties when the Kendall τ is used.

When (X_i, Y_i) are not bivariate normal but ρ exists, according to an approximation of the Kendall τ [Kendall (1949)], using the first fourth-order cumulants and the *bivariate Gram-Charlier series* expansion yield that

$$E(\tau) \approx \frac{2}{\pi} \arcsin(\rho) + \frac{1}{24\pi(1-\rho^2)^{3/2}} \{(\kappa_{40} + \kappa_{04})(3\rho - 2\rho^3) - 4(\kappa_{31} + \kappa_{13}) + 6\rho\kappa_{22}\},$$

where $\kappa_{40} = \mu_{40} - 3$, $\kappa_{31} = \mu_{31} - 3\rho$, $\kappa_{22} = \mu_{22} - 2\rho^2 - 1$. If under some certain conditions that κ_{31} and κ_{13} have a monotonic relationship with ρ and when $\rho = 0$, $\kappa_{31} = 0$ and $\kappa_{13} = 0$, intuitively $E(\tau) = 0$ approximately when $\rho = 0$, and if $|\rho| > c_1$, then there may exist c_2 such that $|E(\tau)| > c_2$. This means that the Kendall' τ based method may enjoy similar properties as the SIS enjoys without strong conditions.

2.2. *Rank correlation screening.* We start our procedure with the linear model as

$$(2.2) \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ is an n -vector of response, $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ is an $n \times p$ random design matrix with independent and identically distributed

$\mathbf{X}_1, \dots, \mathbf{X}_n, \boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is a p -vector of parameters and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ is an n -vector of i.i.d. random errors independent of \mathbf{X} .

To motivate our approach, we briefly review the SIS first. Let

$$(2.3) \quad \boldsymbol{\omega} = (\omega_1, \dots, \omega_p)^T = \mathbf{X}^T \mathbf{Y},$$

where each column of the $n \times p$ design matrix \mathbf{X} has been standardized with mean zero and variance one. Then, for any given $d_n < n$, take the selected submodel to be

$$\widehat{\mathcal{M}}_{d_n} = \{1 \leq j \leq p : |\omega_j| \text{ is among the first } d_n \text{ largest of all}\}.$$

This reduces the full model of size $p \gg n$ to a submodel with the size d_n . By appropriately choosing d_n , all significant predictors can be selected into the submodel indexed by $\widehat{\mathcal{M}}_{d_n}$ with probability tending to 1; see Fan and Lv (2008).

Similar to Li, Peng and Zhu (2011), let $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_p)^T$ be a p -vector each being

$$(2.4) \quad \omega_k = \frac{1}{n(n-1)} \sum_{i \neq j}^n I(X_{ik} < X_{jk}) I(Y_i < Y_j) - \frac{1}{4}, \quad k = 1, \dots, p,$$

where $I(\cdot)$ denotes the indicator function, and ω_k is the marginal rank correlation coefficient between Y and $\mathbf{X}_{\cdot k}$, which is equal to a quarter of the Kendall τ between Y and $\mathbf{X}_{\cdot k}$. As a U-statistic, ω_k is easy to compute. We can then sort the magnitudes of all the components of $\boldsymbol{\omega} = (\omega_1, \dots, \omega_p)^T$ in a decreasing order and select a submodel

$$(2.5) \quad \widehat{\mathcal{M}}_{d_n} = \{1 \leq k \leq p : |\omega_k| \text{ is among the first } d_n \text{ largest of all}\}$$

or

$$(2.6) \quad \widehat{\mathcal{M}}_{\gamma_n} = \{1 \leq k \leq p : |\omega_k| > \gamma_n\},$$

where d_n or γ_n is a predefined threshold value. Thus, it shrinks the full model indexed $\{1, \dots, p\}$ down to a submodel indexed $\widehat{\mathcal{M}}_{d_n}$ or $\widehat{\mathcal{M}}_{\gamma_n}$ with size $|\widehat{\mathcal{M}}_{d_n}| < n$ or $|\widehat{\mathcal{M}}_{\gamma_n}| < n$. Because of the robustness of the Kendall τ against heavy-tailed distributions, such a screening method is expected to be more robust than the SIS.

Consider a more general model as

$$(2.7) \quad H(Y_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\varepsilon_i, i = 1, \dots, n$, are i.i.d. random errors independent of \mathbf{X}_i with mean zero and an unknown distribution F , and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is a p -vector of parameters, its norm constrained to 1 ($\|\boldsymbol{\beta}\| = 1$) for identifiability. $H(\cdot)$ is an unspecified strictly increasing function. Model (2.7) has been studied extensively in the econometric and bioinformatic literature and is commonly used

to stabilize the variance of the error and to normalize/symmetrize the error distribution. With different forms of H and F , this model generates many different parametric families of models. For example, when H takes the form of a power function and F follows a normal distribution, model (2.7) reduces to the familiar Box–Cox transformation models [Box and Cox (1964), Bickel and Doksum (1981)]. If $H(y) = y$ or $H(y) = \log(y)$, model (2.7) reduces to the additive and multiplicative error models, respectively. More parametric transformation models can be found in the work of Carroll and Ruppert (1988).

For model (2.7), the invariance against any strictly increasing transformation yields that

$$\begin{aligned}
 \omega_k &= \frac{1}{n(n-1)} \sum_{i \neq j}^n I(X_{ik} < X_{jk}) I(Y_i < Y_j) - \frac{1}{4} \\
 (2.8) \quad &= \frac{1}{n(n-1)} \sum_{i \neq j}^n I(X_{ik} < X_{jk}) I(H(Y_i) < H(Y_j)) - \frac{1}{4}
 \end{aligned}$$

for $k = 1, \dots, p$. That is, $\omega_k, k = 1, 2, \dots, p$, can still be applicable for the model with unknown transformation function. Therefore, the RRCS method can also be applied to transformation regression models that establish the nonlinear relationship between the response and predictor variables.

3. Sure screening properties of RRCS. In this section we study the sure screening properties of RRCS for the linear regression model (2.2) and the transformation regression model (2.7). Without loss of generality, let $(Y_1, X_{1k}), (Y_2, X_{2k})$ be the independent copies of (Y, X_k) , where $EY = EX_k = 0$ and $EY^2 = EX_k^2 = 1, k = 1, \dots, p$, and assume that

$$\mathcal{M}_* = \{1 \leq k \leq p : \beta_k \neq 0\}$$

is the true sparse model with nonsparsity size $s_n = |\mathcal{M}_*|$, recalling that $\beta = (\beta_1, \dots, \beta_p)^T$ is the true parameter vector. The complement of \mathcal{M}_* is

$$\mathcal{M}_*^c = \{1 \leq k \leq p : k \notin \mathcal{M}_*\}.$$

Furthermore, for $k = 1, \dots, p$, let $\rho_k = \text{corr}(X_k, Y)$ for model (2.2) and $\rho_k^* = \text{corr}(X_k, H(Y))$ for model (2.7). Recall the definition of $\omega = \{\omega_1, \dots, \omega_p\}^T$ in (2.4) for both (2.2) and (2.7).

The following marginal conditions on the models are needed to ensure the sure screening properties of RRCS.

Marginally symmetric condition and Multi-modal condition: For model (2.2):

(M1) Denote $\Delta Y = Y_1 - Y_2$, then the conditional distribution $F_{\Delta Y | \Delta X_k}(t)$ is symmetric about zero when $k \in \mathcal{M}_*^c$, where $\Delta X_k = X_{1k} - X_{2k}$.

(M2) Denote $\Delta\epsilon_k = Y_1 - Y_2 - \rho_k(X_{1k} - X_{2k})$ and $\Delta X_k = X_{1k} - X_{2k}$, then the conditional distribution $F_{\Delta\epsilon_k|\Delta X_k}(t) = \pi_{0k}F_0(t, \sigma_0^2|\Delta X_k) + (1 - \pi_{0k})F_1(t, \sigma_1^2|\Delta X_k)$ follows a symmetric finite mixture distribution where $F_0(t, \sigma_0^2|\Delta X_k)$ follows a symmetric unimodal distribution with the conditional variance σ_0^2 related to ΔX_k and $F_1(t, \sigma_1^2|\Delta X_k)$ is a symmetric distribution function with the conditional variance σ_1^2 related to ΔX_k when $k \in \mathcal{M}_*$. $\pi_{0k} \geq \pi^*$, where π^* is a given positive constant in $(0, 1]$ for any ΔX_k and any $k \in \mathcal{M}_*$.

For model (2.7):

(M1') Denote $\Delta H(Y) = H(Y_1) - H(Y_2)$, where $H(\cdot)$ is the link function of the transformation regression model (2.7), and $\Delta X_k = X_{1k} - X_{2k}$. The conditional distribution $F_{\Delta H(Y)|\Delta X_k}(t)$ is symmetric about zero when $k \in \mathcal{M}_*$.

(M2') Denote $\Delta\epsilon_k = H(Y_1) - H(Y_2) - \rho_k^*(X_{1k} - X_{2k})$ and $\Delta X_k = X_{1k} - X_{2k}$, where $H(\cdot)$ is the link function of the transformation regression model (2.7), then the conditional distribution $F_{\Delta\epsilon_k|\Delta X_k}(t) = \pi_{0k}F_0(t, \sigma_0^2|\Delta X_k) + (1 - \pi_{0k})F_1(t, \sigma_1^2|\Delta X_k)$ follows a symmetric finite mixture distribution where $F_0(t, \sigma_0^2|\Delta X_k)$ follows a symmetric unimodal distribution with the conditional variance σ_0^2 related to ΔX_k and $F_1(t, \sigma_1^2|\Delta X_k)$ is a symmetric distribution function with the conditional variance σ_1^2 related to ΔX_k when $k \in \mathcal{M}_*$. $\pi_{0k} \geq \pi^*$, where π^* is a given positive constant in $(0, 1]$ for any ΔX_k and any $k \in \mathcal{M}_*$.

REMARK 1. According to the definition and symmetric form of $\Delta Y, \Delta X_k$ and $\Delta\epsilon_k$, the marginally symmetric conditions (M2) and (M2') are very mild. When π^* is small enough, the distribution is close to F_1 which is naturally symmetric and has no stringent constraint.

A special case is that the conditional distribution of $\epsilon_{ik} = Y_i - \rho_k X_{ik}$ or $\epsilon_{ik} = H(Y_i) - \rho_k^* X_{ik}$, given X_{ik} ($i = 1, \dots, n$), is homogeneous (not depending on X_{ik}) with a finite number of modes. Actually, when this condition holds, the conditional distribution of ϵ_{ik} given X_{ik} is identical to the corresponding unconditional marginal distribution. Note that $\Delta\epsilon_k = \epsilon_{1k} - \epsilon_{2k}$. When $\epsilon_{ik}, i = 1, 2$, follows multimodal distribution $F_\epsilon(t)$ with no more than K modes where K is not related to k and n , such a distribution function can be rewritten as a weighted sum of K unimodal distributions $F_i(\cdot)$ as

$$F_\epsilon(t) = \sum_{i=1}^K \pi_i F_i(t),$$

where $\pi_i \geq 0, i = 1, \dots, K$, with $\sum_{i=1}^K \pi_i = 1$. Then it is easy to see that the distribution of $\Delta\epsilon_k = \epsilon_{1k} - \epsilon_{2k}$ has the following form:

$$F_{\Delta\epsilon}(t) = \sum_{i=1}^K \sum_{j=1}^K \pi_i \pi_j F_{ij}^*(t) = \sum_{i=1}^K \pi_i^2 F_{ii}^*(t) + \sum_{i \neq j}^K \pi_i \pi_j F_{ij}^*(t)$$

$$\begin{aligned}
&= \left\{ \sum_{i=1}^K \pi_i^2 \right\} \sum_{i=1}^K \frac{\pi_i^2}{\sum_{i=1}^K \pi_i^2} F_{ii}^*(t) + \left(1 - \sum_{i=1}^K \pi_i^2 \right) \sum_{i \neq j}^K \frac{\pi_i \pi_j}{1 - \sum_{i=1}^K \pi_i^2} F_{ij}^*(t) \\
&\hat{=} \pi_0^* F_0^{**}(t) + (1 - \pi_0^*) F_1^{**}(t),
\end{aligned}$$

where $F_{ij}^*(t), i, j = 1, \dots, K$, are the distributions of the differences of two independent variables, that is, $Z_i - Z_j$ where Z_i follows the distribution of $F_i(t)$ and Z_j follows the distribution of $F_j(t)$, respectively. Because $F_i(t), i = 1, \dots, K$, are unimodal distributions, $F_{ii}^*, i = 1, \dots, K$, are then symmetric unimodal distributions. Hence, $F_0^{**}(t)$ is a symmetric unimodal distribution. It is also easy to see that $F_1^{**}(t)$ is a symmetric multimodal distribution function. On the other hand, $\pi_0^* = \sum_{i=1}^K \pi_i^2 \geq 1/K(\sum_{i=1}^K \pi_i)^2 = 1/K$. As such, (M2) or (M2') is satisfied.

Other than the marginally symmetric conditions, we also need the following regularity conditions:

(C1) As $n \rightarrow +\infty$, the dimensionality of \mathbf{X} satisfies $p = O(\exp(n^\delta))$ for some $\delta \in (0, 1)$, satisfying $\delta + 2\kappa < 1$ for any $\kappa \in (0, \frac{1}{2})$.

(C2) $c_{\mathcal{M}_*} = \min_{k \in \mathcal{M}_*} E|X_{1k}|$ is a positive constant and is free of p .

(C3) The predictors \mathbf{X}_i and the error $\varepsilon_i, i = 1, \dots, n$, are independent of one another.

REMARK 2. Condition (C1) guarantees that for the independence screening method, we can select significant predictors into a working submodel with probability tending to 1. SIS also needs this condition; see Fan and Lv (2008) and Fan and Song (2010). Condition (C2) is a mild technical condition that ensures the sure screening property of the RRCS procedure. It is worth mentioning that we do not need to have a uniform bound for all EX_{1k}^2 . If the size of \mathcal{M}_* goes to infinity with a relatively slow speed, we can relax this condition to $c_{\mathcal{M}_*} > cn^{-\iota}$ for some positive constant c and $\iota \in (0, 1)$ with a suitable choice of the threshold γ_n . Precisely, γ_n can be chosen as $c'n^{-\kappa-\iota}$ for some positive constant c' where κ satisfies $2\kappa + 2\iota < 1$. From Theorem 1 below, we can see that $|E(\omega_k)| > cn^{-\kappa-\iota}$ for $k \in \mathcal{M}_*$. To ensure the sure screening properties, (C1) needs to be changed to $\delta + 2\kappa + 2\iota < 1$.

THEOREM 1. *Under the regularity condition (C2) and the marginal symmetric conditions (M1) and (M2) for model (2.2), we have the following:*

- (i) $E(\omega_k) = 0$ if and only if $\rho_k = 0$.
- (ii) If $|\rho_k| > c_1 n^{-\kappa}$ for $k \in \mathcal{M}_*$ with a positive constant $c_1 > 0$, then there exists a positive constant c_2 such that $\min_{k \in \mathcal{M}_*} |E(\omega_k)| > c_2 n^{-\kappa}$.

For model (2.7), replacing conditions (M1) and (M2) with (M1') and (M2'), then:

- (i') $E(\omega_k) = 0$ if and only if $\rho_k^* = 0$.

(ii') If $|\rho_k^*| > c_1 n^{-\kappa}$ for $k \in \mathcal{M}_*$ with a positive constant $c_1 > 0$, then there exists a positive constant c_2 such that $\min_{k \in \mathcal{M}_*} |\mathbb{E}(\omega_k)| > c_2 n^{-\kappa}$.

REMARK 3. As Fan and Song (2010) mentioned, the marginally symmetric condition (M1) is weaker than the partial orthogonality condition assumed by Huang, Horowitz and Ma (2008), that is, $\{X_k, k \in \mathcal{M}_*^c\}$ is independent of $\{X_k, k \in \mathcal{M}_*\}$, which can lead to the model selection consistency for the linear model. Our results, together with the following Theorem 2, indicate that under weaker conditions, consistency can also be achieved even for transformation regression models. Furthermore, as in the discussion of Fan and Song (2010), a necessary condition for the sure screening is that the significant predictors X_k with $\beta_k \neq 0$ are correlated with the response in the sense that $\rho_k \neq 0$. The result (i) of Theorem 1 also shows that when the Kendall τ is used, this property can be held, which suggests that the insignificant predictors in \mathcal{M}_*^c can be detected from $\mathbb{E}(\omega_k)$ at the population level. Result (ii) indicates that under marginally symmetric conditions, a suitable threshold γ_n can entail the sure screening in the sense of

$$\min_{k \in \mathcal{M}_*} |\mathbb{E}(\omega_k)| \geq \gamma_n, \quad \max_{k \in \mathcal{M}_*^c} |\mathbb{E}(\omega_k)| = 0.$$

REMARK 4. As a by-product, Theorem 1 reveals the relationship between the Pearson correlation and the Kendall τ under general conditions, especially the multi-modal conditions (M2) or (M2') which in itself is of interest. However, either condition (M2) or (M2') is a sufficient condition to guarantee that the Kendall τ has either the property (ii) or (ii') of Theorem 1, and then has the sure screening property. As in the discussion in Section 2.1, following the high order *bivariate Gram-Charlier series expansion* to approximate the joint distribution of (X_i, Y_i) , under certain conditions such as either the condition or sub-Gaussian tail condition, we could also obtain similar results of Theorem 1. It would involve some high order of moments or cumulants. However, as shown in Theorem 1, either the multi-modal condition (M2) or (M2') is to ensure the robust properties of the proposed RRCS, and depicts those properties more clearly. Furthermore, we will show in the proposition below that the bivariate normal copula family also makes another sufficient condition for the following Theorem 1 to hold.

Bivariate normal copula family based marginal condition: We give another sufficient condition for (X_i, Y_i) for the results of Theorem 1 to hold. Consider the bivariate normal copula family which is defined as

$$C_\theta(u_1, u_2) = \Phi_\theta(\Phi^{-1}(u_1), \Phi^{-1}(u_2)), \quad 0 \leq u_1, u_2 \leq 1,$$

where Φ_θ is a bivariate standard normal distribution function with mean zero, variance one and correlation θ , Φ is the one-dimensional standard normal distribution function. Let \mathcal{F} denote the collection of all distribution

functions on \mathbb{R} . We then define the bivariate distribution family \mathcal{P} as

$$\mathcal{P} = \{C_\theta(F_X(x), F_Y(y)), (x, y) \in \mathbb{R}^2, F_X \in \mathcal{F}, F_Y \in \mathcal{F}\}.$$

Copula now is a popular tool to study the dependence among multivariate random variables. For details, see Nelsen (2006). The normal copula family is an important copula family in practice. Particularly, the bivariate normal copula family can be used to approximate most of the distributions of bivariate continuous or discrete random vectors, for example, see Cario and Nelson (1997), Ghosh and Henderson (2003), Pitt, Chan and Kohn (2006) and Channouf and L'Ecuyer (2009).

Based on the results of Klaassen and Wellner (1997) and the monotonic relationship between the Kendall τ and the Pearson correlation, the multimodality can be replaced by the above copula distribution family. A proposition is stated below.

PROPOSITION 1. *Under the marginal symmetric condition (M1) for model (2.2), we have the following:*

- (i) $E(\omega_k) = 0$ if and only if $\rho_k = 0$.
- (ii) If $|\rho_k| > c_1 n^{-\kappa}$ with a positive constant $c_1 > 0$ and the joint distribution $F(x, y)$ of (X_k, Y) is in \mathcal{P} , for $k \in \mathcal{M}_*$, then there exists a positive constant c_2 such that $\min_{k \in \mathcal{M}_*} |E(\omega_k)| > c_2 n^{-\kappa}$.

For model (2.7), replacing condition (M1) with (M1'), then:

- (i') $E(\omega_k) = 0$ if and only if $\rho_k^* = 0$.
- (ii') If $|\rho_k^*| > c_1 n^{-\kappa}$ with a positive constant $c_1 > 0$ and the joint distribution $F(x, y)$ of (X_k, Y) is in \mathcal{P} for $k \in \mathcal{M}_*$, then there exists a positive constant c_2 such that $\min_{k \in \mathcal{M}_*} |E(\omega_k)| > c_2 n^{-\kappa}$.

REMARK 5. If the joint distribution of (X, Y) is in \mathcal{P} with the formula $F(X, Y) = C_\theta(F_X(X), F_Y(Y))$, the results of Klaassen and Wellner (1997) suggested that $|\theta|$ equals the maximum correlation coefficient between X and Y . As shown in the proof of the proposition, when we replace ρ by θ in the proposition, the results continue to hold. Hence, this proposition provides a bridge between our method and the generalized correlation proposed by Hall and Miller (2009) because, according to their definitions, the generalized correlation coefficient is an approximation of the maximum correlation coefficient.

Sure screening property of RRCS: Based on Theorem 1 or Proposition 1, the sure screening property and model selection consistency of RRCS are stated in the following results.

THEOREM 2. *Under the conditions (C1)–(C3), and the conditions of Theorem 1 or Proposition 1 corresponding to either model (2.2) or model (2.7), for some $0 < \kappa < 1/2$ and $c_3 > 0$, there exists a positive constant*

$c_4 > 0$ such that

$$\mathbb{P}\left(\max_{1 \leq j \leq p} |\omega_j - E(\omega_j)| \geq c_3 n^{-\kappa}\right) \leq p\{\exp(-c_4 n^{1-2\kappa})\}.$$

Furthermore, by taking $\gamma_n = c_5 n^{-\kappa}$ with $c_5 \leq c_2/2$, if $|\rho_k| > c_1 n^{-\kappa}$ for $j \in \mathcal{M}_*$, we have

$$\mathbb{P}(\mathcal{M}_* \subset \widehat{\mathcal{M}}_{\gamma_n}) \geq 1 - 2|\mathcal{M}_*|\{\exp(-c_4 n^{1-2\kappa})\}.$$

REMARK 6. Theorem 2 shows that RRCS can handle the NP-dimensionality problem for linear and semiparametric transformation regression models. It also permits $\log p = o(n^{1-2\kappa})$, which is identical to that in Fan and Lv (2008) for the linear model and is faster than $\log p = o(n^{(1-2\kappa)/A})$ with $A = \max(\alpha + 4, 3\alpha + 2)$ for some positive α in Fan and Song (2010) when the likelihood ratio screening is used.

REMARK 7. It is obvious when the joint distribution of (\mathbf{X}_i^T, Y_i) follows a multivariate normal distribution, conditions (M1) and (M2) are automatically valid. The results of sure screening properties are equivalent to those of Fan and Lv (2008) under weaker conditions. This is because of the definition of the rank correlation Kendall τ and its monotonic relationship with the Pearson correlation as in the discussion in Section 2. The Kendall τ can be regarded as a U-statistic and uses the indicator function as the link function. As the indicator function is a bounded function, the exponential U-statistic inequality can be used to directly control the tail of the rank correlation Kendall τ rather than those of \mathbf{X}_i and Y_i .

Under the conditions of Proposition 1, following similar steps, the same results of Theorem 2 and the following Theorem 3 can be obtained without any difficulties. Thus, we only present the relevant results without the detailed technical proofs.

The following theorem states that the size of $\widehat{\mathcal{M}}_{\gamma_n}$ can be controlled by the RRCS procedure.

THEOREM 3. Under the conditions (C1)–(C3), and conditions of Theorem 1 or Proposition 1 for model (2.2), when $|\rho_k| > c_1 n^{-\kappa}$ for some positive constant c_1 uniformly in $k \in \mathcal{M}_*$, for any $\gamma_n = c_5 n^{-\kappa}$ there exists a constant $c_6 > 0$ such that

$$(3.1) \quad \mathbb{P}(|\widehat{\mathcal{M}}_{\gamma_n}| \leq O\{n^{2\kappa} \lambda_{\max}(\Sigma)\}) \geq 1 - p\{\exp(-c_6 n^{1-2\kappa})\},$$

where $\Sigma = \text{Cov}(\mathbf{X}_i)$ and $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$. For model (2.7) in addition to conditions (C1)–(C3) and the marginal symmetric conditions (M1') and (M2'), when $|\rho_k^*| > c_1 n^{-\kappa}$ for some positive constant c_1 uniformly in $k \in \mathcal{M}_*$ and $\text{Var}(H(Y)) = O(1)$, for $\gamma_n = c_5 n^{-\kappa}$ there exists a constant $c_6 > 0$ such that the above inequality (3.1) holds.

REMARK 8. Compared with Theorem 5 of Fan and Song (2010), the conditions of Theorem 3 are much weaker and the obtained inequalities are much simpler in form although the rates are similar. The number of selected predictors is of the order $\|\Sigma\beta\|/\gamma_n^2$, which is bounded by $O\{n^{2\kappa}\lambda_{\max}(\Sigma)\}$ when $\text{Var}(H(Y)) = O(1)$. Hence, when $\lambda_{\max}(\Sigma) = O(n^\tau)$, the size of the selected predictors is of the order $O(n^{2\kappa+\tau})$, which can be smaller than n when $2\kappa + \tau < 1$.

From Theorems 1–3, the rank correlation has sure screening properties and model selection consistency. However, it is also obvious that it does not sufficiently use all of the information from data, particularly the correlations of predictors. Hence, as most of the other sure screening methods, the rank sure screening can be only regarded as an initial model selection reducing the ultra-high dimension down to a dimension smaller than the sample size n without losing any important significant predictor variables. As the numerical results in Section 5 and the discussion of Fan and Lv (2008) show, the correlation of predictors could seriously affect the sure screening results, and thus more subtle sure screening methods, such as Iterative Sure Independence Screening (ISIS) [Fan and Lv (2008)], are in need.

4. IRRCS: Iterative robust rank correlation screening.

4.1. *IRRCs*. With RRCS, the dimension can be brought down to a value smaller than the sample size with a probability tending to one. Thus, we can work on a smaller submodel. However, in most situations, RRCS can be only regarded as a crude model selection method, and the resulting model may still contain many superfluous predictors. It is partly because strong correlation always exists between predictors when too many predictors are involved [see Fan and Lv (2008)], and the basic sure screening methods do not use this correlation information. We also face some other issues. First, in modeling high dimensional data, it is often a challenge to determine outliers. High dimensionality also increases the likelihood of extreme values of predictors. Second, even when the model dimension is smaller than the sample size, the design matrix may still be near singular when strong correlation exists between predictors. Third, the usual normal or sub-Gaussian distributional assumption on predictors/errors is not easy to substantiate. Fourth, it is also an unfortunate fact that the RRCS procedure may break down if a predictor is marginally unrelated but jointly related with the response, or if a predictor is jointly unrelated with the response but has higher marginal correlation with the response than some significant predictors. To deal with these issues, we develop a robust iterative RRCS (IRRCs) that is motivated by the concept of Iterative Sure Independence Screening (ISIS) in Fan and Lv (2008).

To this end, we first briefly describe a penalized smoothing maximum rank correlation estimator (PSMRC) suggested by Lin and Peng (2013). This estimation approach is applied to simultaneously further select and estimate a final working submodel through working on β .

For model (2.7), the monotonicity of H and the independence of \mathbf{X} and ϵ ensure that

$$\mathbb{P}(Y_i \geq Y_j | \mathbf{X}_i, \mathbf{X}_j) \geq \mathbb{P}(Y_i \leq Y_j | \mathbf{X}_i, \mathbf{X}_j) \quad \text{whenever } \mathbf{X}_i^T \beta \geq \mathbf{X}_j^T \beta.$$

Hence, β can be estimated by maximizing

$$(4.1) \quad G_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i > Y_j) I(\mathbf{X}_i^T \beta > \mathbf{X}_j^T \beta).$$

It is easy to see that $G_n(\beta)$ is another version of the Kendall τ between Y_i and $\mathbf{X}_i^T \beta$. The maximum rank correlation [MRC; Han (1987)] estimator $\hat{\beta}_n$ can be applied to estimate β . When p is fixed, the $n^{1/2}$ -consistency and the asymptotic normality of $\hat{\beta}_n$ have been derived. However, because $G_n(\beta)$ is not a smooth function, the Newton–Raphson algorithm cannot be used directly, and the optimization of $G_n(\beta)$ requires an intensive search at heavy computational cost. We then consider PSMRC as follows. Define

$$(4.2) \quad L_n(\beta) = S_n(\beta) - \sum_{j=1}^d p_{\lambda_n}(|\beta_j|)$$

and

$$(4.3) \quad S_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i > Y_j) \Phi((\mathbf{X}_i - \mathbf{X}_j)^T \beta / h),$$

where $\Phi(\cdot)$ is the standard normal distribution function, a smooth function for the purpose of reducing computational burden, h is a small positive constant, and $p_\lambda(|\cdot|)$ is a penalty function of L_1 type such as that in LASSO, SCAD or MCP. It is easy to see if $h \rightarrow 0$, $\Phi((\mathbf{X}_i - \mathbf{X}_j)^T \beta / h) \rightarrow I(\mathbf{X}_i^T \beta > \mathbf{X}_j^T \beta)$. As $L_n(\beta)$ is a smoothing function of β , traditional optimal methods, such as the Newton–Raphson algorithm or newly developed LARS [Efron et al. (2004)] and LLA [Zou and Li (2008)], can be used to obtain the maximizer of $L_n(\beta)$ to simultaneously achieve the selection and estimation of β . For model (2.2), the problem is easier and we do not repeatedly describe the estimation for it.

Next, we introduce our intuitive idea for the proposed IRRCS for the transformation regression model. Such an idea can be also applied to the linear model since it is a special transformation regression model. In fact, given the i.i.d. sequences Y_i and $\mathbf{X}_i^T \beta, i = 1, \dots, n$, define $Y_{ij}^* = I(Y_i < Y_j)$ and $\mathbf{X}_{ij}^*(\beta) = I(\mathbf{X}_i \beta < \mathbf{X}_j \beta)$. Then the Pearson correlation between Y_{ij}^* and

$\mathbf{X}_{ij}^*(\boldsymbol{\beta})$ is the rank correlation Kendall τ between Y_i and $\mathbf{X}_i\boldsymbol{\beta}$. According to the idea of the maximum rank correlation [MRC; Han (1987)] estimator, the estimate of $\boldsymbol{\beta}$ for the transformation regression model just maximizes the Pearson correlation between Y_{ij}^* and $\mathbf{X}_{ij}^*(\boldsymbol{\beta})$ or the rank correlation Kendall τ between Y_i and $\mathbf{X}_i\boldsymbol{\beta}$. If we do not care about the norm of $\boldsymbol{\beta}$, the least squares estimate of $\boldsymbol{\beta}$ in the linear model just maximizes the Pearson correlation between Y_i and $\mathbf{X}_i^T\boldsymbol{\beta}$. If we regard the transformation model as the following special linear model:

$$Y_{ij}^* = \mathbf{X}_{ij}^*(\boldsymbol{\beta}) + \varepsilon_{ij},$$

where $\varepsilon_{ij} = I(\varepsilon_i < \varepsilon_j)$. Then it is easy to see that MRC for the transformation model and the least squares estimate for the linear model are based on a similar principle and, hence, the idea of Iterative Sure Independence Screening (ISIS) for the linear model in Fan and Lv (2008) can be used for the transformation model. Based on this intuitive insight, our proposed IRRCS procedure is as follows:

Step 1. First the RRCS procedure is used to reduce the original dimension to a value $\lceil n/\log n \rceil$ smaller than n . Then, based on the joint information from the $\lceil n/\log n \rceil$ predictors that survive after the RRCS, we select a subset of d_1 predictors $\mathcal{M}_1 = \{X_{i_1}, \dots, X_{i_{d_1}}\}$ by a model selection method such as the nonconcave penalized M-estimation proposed by Li, Peng and Zhu (2011) for model (2.2) and the penalized smoothing maximum correlation estimator [Lin and Peng (2013)] for model (2.7).

Step 2. Let $\mathbf{X}_{i,\mathcal{M}_1} = (X_{i_1}, \dots, X_{i_{d_1}})^T$ be the $d_1 \times 1$ vector selected in step 1, and $l = 1, \dots, p - d_1$.

- For model (2.2), define $Y_i^* = Y_i - \mathbf{X}_{i,\mathcal{M}_1}^T \hat{\boldsymbol{\beta}}_{\mathcal{M}_1}$, then the Kendall τ values for the remaining $p - d_1$ predictors are calculated as follows:

$$\omega_l = \frac{1}{n(n-1)} \sum_{j \neq i}^n I(Y_i^* < Y_j^*) I(X_{il} < X_{jl}) - \frac{1}{4},$$

where $\hat{\boldsymbol{\beta}}_{\mathcal{M}_1}$ is a vector estimator of the d_1 nonzero coefficients that are estimated by the nonconcave penalized M-estimate method in Li, Peng and Zhu (2011). Sort the $p - d_1$ values of the $|\omega_l|$ again and select another subset of $\lceil n/\log n \rceil$ predictors from $\mathcal{M} - \mathcal{M}_1$.

- For model (2.7), define $I(Y_i^*, Y_j^*) = I(Y_i, Y_j) - I(\mathbf{X}_{i,\mathcal{M}_1}^T \hat{\boldsymbol{\beta}}_{\mathcal{M}_1} < \mathbf{X}_{j,\mathcal{M}_1}^T \hat{\boldsymbol{\beta}}_{\mathcal{M}_1})$ where $I(Y_i, Y_j) = I(Y_i < Y_j)$ where $\hat{\boldsymbol{\beta}}_{\mathcal{M}_1}$ is an estimator of the d_1 nonzero coefficients, which are estimated with the penalized smoothing maximum correlation estimator of Lin and Peng (2013). Then, compute the Kendall τ through the remaining $p - d_1$ predictors as

$$\omega_l = \frac{1}{n(n-1)} \sum_{j \neq i}^n I(Y_i^*, Y_j^*) I(X_{il} < X_{jl}) - \frac{1}{4},$$

and sort the $p - d_1$ values of the $|\omega_l|$'s again and select a subset of $[n/\log n]$ predictors as in step 1.

Step 3. Replace Y_i by Y_i^* in (2.2) and $I(Y_i, Y_j)$ with $I(Y_i^*, Y_j^*)$ in (4.2), and select a subset of d_2 predictors $\mathcal{M}_2 = \{X_{i_1}, \dots, X_{i_{d_2}}\}$ from the joint information of the $[n/\log n]$ predictors that survived in step 2 as in step 1.

Step 4. Iterate steps 2 and 3 until k disjoint subsets $\mathcal{M}_1, \dots, \mathcal{M}_k$ are obtained whose union $\mathcal{M} = \bigcup_{i=1}^k \mathcal{M}_i$ has a size d less than sample size n . In the implementation, we can choose, for example, the largest k such that $|\mathcal{M}| < n$.

4.2. *Discussion on RRCS for generalized linear and single-index models.* Consider the generalized linear model

$$(4.4) \quad f_Y(y, \theta) = \exp\{y\theta - b(\theta) + c(y)\}$$

for known functions $b(\cdot)$ and $c(\cdot)$ and unknown function θ , where the dispersion parameter is not considered as the mean regression modeled. The function θ is usually called canonical or a natural parameter, and the following structure of the generalized linear model is often considered:

$$(4.5) \quad E(Y|\mathbf{X} = \mathbf{x}) = b'(\theta(\mathbf{x})) = g^{-1}\left(\sum_{j=0}^p \beta_j x_j\right),$$

where $\mathbf{x} = (x_0, \dots, x_p)^T$ is a $(p+1)$ -dimensional predictor, $x_0 = 1$ represents the intercept, and $\theta(\mathbf{x}) = \sum_{j=0}^p \beta_j x_j$. In this case, $g(\cdot)$ should be a strictly increasing function. Thus, we may use ω of (2.8) with function g^{-1} to rank the importance of the predictors. Although the idea seems straightforward, the technical details are not easily handled, and we leave them to further study. In the simulations, we examine its performance; see the details in Section 5. In addition, after reducing the dimension, we consider estimating the parameters in the working submodel. Again, we can also see that

$$\mathbb{P}(Y_i \geq Y_j | \mathbf{X}_i, \mathbf{X}_j) \geq \mathbb{P}(Y_i \leq Y_j | \mathbf{X}_i, \mathbf{X}_j) \quad \text{whenever } \mathbf{X}_i^T \boldsymbol{\beta} \geq \mathbf{X}_j^T \boldsymbol{\beta}.$$

Hence, Han's (1987) MRC estimator can be used. Fan and Song (2010) applied the idea of SIS to (4.4) with NP-dimensionality, and used the maximum marginal likelihood estimator (MMLE). They showed that the MMLE $\beta_j^M = 0$ if and only if $\text{Cov}(b'(\mathbf{X}^T \boldsymbol{\beta}), X_j) = \text{Cov}(Y, X_j) = 0$. That is, MMLE is equivalent to the Pearson correlation in a certain sense when SIS is applied.

A further generalization is with unknown canonical link function $g(\cdot)$. In this case, the generalized linear model can be regarded as a special single index model with a strictly increasing restriction as the link function $b'(\cdot)$ or $g(\cdot)$. Based on the discussion in Section 2, we can also use the Kendall τ based method to select predictors and PSMRC to estimate the parameters. The selection and estimation could be more robust than with the MMLE based SIS.

5. Numerical studies and application.

5.1. *Simulations.* In the first 4 examples, we compare the performance of the five methods: SIS, ISIS, RRCS, IRRCS, and the generalized correlation rank method (gcorr) proposed by Hall and Miller (2009) by computing the frequencies with which the selected models include all of the variables in the true model, that is, their ability to correctly screen unimportant variables. The simulation examples cover the linear models used by Fan and Lv (2008), the transformation models used by Lin and Peng (2013), the Box–Cox transformation model used by Hall and Miller (2009), and the generalized linear models used by Fan and Song (2010). We also use a “semi-real” example as Example 5, in which a part of the data are from a real data set and the other part of the data are artificial. The difference from the other examples is that this data set contains categorical data.

EXAMPLE 1. Consider the following linear model:

$$(5.1) \quad Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\boldsymbol{\beta} = (5, 5, 5, 0, \dots, 0)^T$, $\mathbf{X}_i = (X_{1i}, \dots, X_{pi})^T$ is a p -dimensional predictor and the noise ε_i is independent of the predictors, and is generated from three different distributions: the standard normal, the standard normal with 10% of the outliers following the Cauchy distribution and the standard t distribution with three degrees of freedom. The first $k = 3$ predictors are significant, but the others are not. \mathbf{X}_i are generated from a multivariate normal distribution $N(0, \Sigma)$ with entries of $\Sigma = (\sigma_{ij})_{p \times p}$ being $\sigma_{ii} = 1, i = 1, \dots, p$, and $\sigma_{ij} = \rho, i \neq j$. For some combinations with $p = 100, 1000$, $n = 20, 50, 70$ and $\rho = 0, 0.1, 0.5, 0.9$, the experiment is repeated 200 times.

As different methods may select a working model with different sizes, to ensure a fair comparison, we select the same size of $n - 1$ predictors using the four methods. Then we check their selection accuracy in including the true model $\{X_1, X_2, X_3\}$. The details of ISIS can be found in Section 4 of Fan and Lv (2008). In Table 1, we report the proportions of predictors containing the true model selected by RRCS, SIS, IRRCS and ISIS.

From Table 1, we can draw the following conclusions:

(1) When noise ε is drawn from the standard normal, SIS and ISIS perform better than RRCS and IRRCS according to higher proportions of predictors containing the true model selected. The difference becomes smaller with a larger sample size and smaller ρ . ISIS and IRRCS can greatly improve the performance of SIS and RRCS. IRRCS can outperform ISIS.

(2) When $\rho = 0.5$ or 0.9 , SIS and RRCS perform worse than in the cases with $\rho = 0$ or 0.1 . This coincides with our intuition that high collinearity deteriorates the performance of SIS and RRCS.

(3) It is also worth mentioning that even when there are outliers or the heavy-tailed errors, RRCS is not necessarily better than SIS. This is an in-

TABLE 1
Example 1: the proportion of predictors containing the true model $\{X_1, X_2, X_3\}$ selected by RRCS, SIS, IRRCS and ISIS

(p, n)	$\varepsilon \sim$	$N(0, 1)$				$N(0, 1)$ with 10% outliers				$t(3)$			
	Method	$\rho = 0$	0.1	0.5	0.9	0	0.1	0.5	0.9	0	0.1	0.5	0.9
(100, 20)	RRCS	0.765	0.745	0.605	0.405	0.840	0.835	0.730	0.640	0.850	0.840	0.765	0.520
	SIS	0.835	0.875	0.725	0.650	0.810	0.845	0.705	0.590	0.775	0.805	0.600	0.315
	IRRCS	0.840	0.905	0.865	0.915	0.995	0.980	0.960	0.895	0.995	1	0.995	0.930
	ISIS	1	1	0.985	0.985	0.885	0.850	0.855	0.845	0.895	0.910	0.865	0.845
(100, 50)	RRCS	1	1	1	0.985	0.980	0.960	0.970	0.930	1	0.995	0.980	0.965
	SIS	1	1	1	1	0.960	0.950	0.970	0.915	0.965	0.970	0.960	0.920
	IRRCS	1	1	1	1	1	1	1	0.970	1	1	1	0.990
	ISIS	1	1	1	1	0.985	0.975	0.975	0.945	1	1	0.980	0.955
(1000, 20)	RRCS	0.145	0.165	0.060	0.235	0.245	0.250	0.155	0.110	0.245	0.325	0.225	0.150
	SIS	0.255	0.285	0.110	0.140	0.250	0.265	0.125	0.110	0.300	0.270	0.220	0.110
	IRRCS	0.475	0.460	0.480	0.345	0.825	0.840	0.620	0.465	0.860	0.895	0.680	0.580
	ISIS	0.835	0.865	0.715	0.530	0.795	0.840	0.650	0.430	0.805	0.855	0.630	0.460
(1000, 50)	RRCS	0.990	0.970	0.825	0.570	0.945	0.990	0.755	0.555	1	0.990	0.930	0.750
	SIS	1	0.985	0.935	0.835	0.950	0.985	0.845	0.655	0.985	0.985	0.810	0.620
	IRRCS	1	1	0.990	0.995	0.980	0.995	0.950	0.865	1	1	1	0.985
	ISIS	1	1	1	0.995	0.955	0.990	0.940	0.850	1	0.990	0.935	0.850
(1000, 70)	RRCS	1	1	0.990	0.870	0.945	0.990	0.965	0.835	1	1	0.980	0.860
	SIS	1	1	0.990	0.965	0.960	0.950	0.925	0.875	1	0.990	0.950	0.850
	IRRCS	1	1	1	1	1	1	0.975	0.965	1	1	1	1
	ISIS	1	1	1	1	0.970	0.960	0.950	0.940	1	1	0.980	0.960

interesting observation. However, when we note the signal-to-noise ratio, we may have an answer. Regardless of outliers, model (5.1) has a large signal-to-noise ratio by taking the nonzero coefficients $(\beta_1, \beta_2, \beta_3) = (5, 5, 5)$. This means that the impact of the outliers on the results is relatively small and RRCS, a nonparametric method, may not be able to show its advantages. We have also tried other simulations with smaller signal-to-noise ratios or larger percentages of outliers. When data has larger percentages of outliers, the performance of RRCS was better than SIS. Especially when iteration is used, IRRCS can outperform the corresponding ISIS even in the case without outliers. When the data has smaller signal-to-noise ratios, for example, $(\beta_1, \beta_2, \beta_3, 0, \dots, 0) = (1, 2/3, 1/3, 0, \dots, 0)$, though the performance of SIS and RRCS are comparable and encouraging, all of the results are not as good as the results of SIS and RRCS in Table 1. This is reasonable, as for all variable selection methods, the phenomenon is the same: when the signal-to-noise ratio becomes smaller, selecting significant predictors gets more difficult.

(4) When the data are contaminated with 10% outliers or are generated from the $t(3)$ distribution, the IRRCS performs better than the ISIS procedure because we use the nonconcave penalized M-estimation in the iterative step for IRRCS.

EXAMPLE 2. Consider Example III in Section 4.2.3 of Fan and Lv (2008) with the underlying model, for $\mathbf{X} = (X_1, \dots, X_p)^T$,

$$(5.2) \quad Y = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + X_5 + \varepsilon,$$

except that X_1, X_2, X_3 and noise ε are distributed identical to those in Example 1 above. For model (5.2), $X_4 \sim N(0, 1)$ has correlation coefficient $\sqrt{\rho}$ with all other $p - 1$ variables, whereas $X_5 \sim N(0, 1)$ is uncorrelated with all the other $p - 1$ variables. X_5 has the same proportion of contributions to the response as ε does, and has an even weaker marginal correlation with Y than X_6, \dots, X_p do. We take $\rho = 0.5$ for simplicity. We generate 200 data sets for this model and report in Table 2 the proportion of RRCS, SIS, IRRCS and ISIS that can include the true model.

The results in Table 2 allow us to draw different conclusions than those from Example 1. Even in the case without outliers or the heavy-tailed errors, SIS and ISIS are not definitely better than RRCS and IRRCS, respectively, whereas in the cases with outliers or heavy-tailed errors there is no exception for IRRCS to work well and better than ISIS. However, the small proportions of RRCS and SIS show their bad performance.

EXAMPLE 3. Consider the following generalized Box–Cox transformation model:

$$(5.3) \quad H(Y_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

TABLE 2
For Example 2: the proportion of RRCS, SIS, IRRCS and ISIS that include the true model $\{X_1, X_2, X_3, X_4, X_5\}$ ($\rho = 0.5$)

p	$\varepsilon \sim$ Method	$N(0, 1)$			$N(0, 1)$ with 10% outliers			$t(3)$		
		$n = 20$	$n = 50$	$n = 70$	$n = 20$	$n = 50$	$n = 70$	$n = 20$	$n = 50$	$n = 70$
100	RRCS	0	0.305	0.595	0	0.220	0.575	0	0.305	0.575
	SIS	0	0.285	0.535	0	0.195	0.525	0	0.240	0.535
	IRRCS	0	0.500	0.820	0	0.495	0.815	0	0.530	0.805
	ISIS	0	0.465	0.855	0	0.415	0.805	0	0.405	0.775
1000	RRCS	0	0	0	0	0	0	0	0	0
	SIS	0	0	0	0	0	0	0	0	0
	IRRCS	0	0.035	0.085	0	0.030	0.055	0	0.030	0.085
	ISIS	0	0.045	0.090	0	0.015	0.035	0	0	0.020

where the transformation functions are unknown. In the simulations, we consider the following forms:

- Box–Cox transformation, $\frac{|Y|^\lambda \operatorname{sgn}(Y) - 1}{\lambda}$, where $\lambda = 0.25, 0.5, 0.75$;
- Logarithm transformation function, $H(Y) = \log Y$.

The linear regression model and the logarithm transformation model are special cases of the generalized Box–Cox transformation model with $\lambda = 1$ and $\lambda = 0$, respectively. Again, noise ε_i follows the distributions as those in the above examples, $\beta = (3, 1.5, 2, 0, \dots, 0)^T$ and $\beta/\|\beta\| = (0.7682, 0.3841, 0.5121, 0, \dots, 0)^T$ is a $p \times 1$ vector, and a sample of $(X_1, \dots, X_p)^T$ with size n is generated from a multivariate normal distribution $N(0, \Sigma)$ whose covariance matrix $\Sigma = (\sigma_{ij})_{p \times p}$ has entries $\sigma_{ii} = 1, i = 1, \dots, p$, and $\sigma_{ij} = \rho, i \neq j$. The replication time is again 200, and $p = 100, 1000$, $n = 20, 50, 70$ and $\rho = 0, 0.1, 0.5, 0.9$, respectively. We also compare the proposed method with the generalized correlation rank method (gcorr) proposed by Hall and Miller (2009) for the logarithm transformation model (the results for the Box–Cox transformation model are similar).

From Tables 3 and 4, we can see clearly that without exception RRCS outperforms SIS and gcorr significantly and IRRCS can greatly improve the performance of RRCS.

EXAMPLE 4 (Logistic regression). In this example, the data $(\mathbf{X}_1^T, Y_1), \dots, (\mathbf{X}_n^T, Y_n)$ are independent copies of a pair (\mathbf{X}^T, Y) , where the conditional distribution of the response Y given X is a binomial distribution with

$$(5.4) \quad \log \left(\frac{p(\mathbf{X})}{1 - p(\mathbf{X})} \right) = \mathbf{X}^T \beta.$$

TABLE 3
Proportion of SIS, RRCS and IRRCS that include the true model for the Box-Cox transformation model $\{X_1, X_2, X_3\}$

(p, n)	λ	$\varepsilon \sim$	$N(0, 1)$				$N(0, 1)$ with 10% outliers				$t(3)$			
		Method	$\rho = 0$	0.1	0.5	0.9	0	0.1	0.5	0.9	0	0.1	0.5	0.9
(100, 20)	0.75	SIS	0.415	0.470	0.190	0.030	0.380	0.435	0.170	0.005	0.420	0.525	0.355	0.200
		RRCS	0.440	0.525	0.400	0.225	0.430	0.510	0.370	0.220	0.525	0.555	0.450	0.220
		IRRCS	0.985	0.975	0.975	0.850	0.940	0.910	0.875	0.755	0.960	0.945	0.925	0.840
	0.5	SIS	0.320	0.390	0.155	0.005	0.265	0.345	0.160	0.005	0.360	0.490	0.325	0.090
		RRCS	0.435	0.525	0.400	0.225	0.450	0.510	0.390	0.195	0.590	0.545	0.355	0.225
		IRRCS	0.985	0.970	0.945	0.860	0.900	0.890	0.885	0.745	0.935	0.920	0.910	0.815
	0.25	SIS	0.150	0.195	0.090	0.0025	0.145	0.155	0.085	0.0015	0.190	0.225	0.175	0.005
		RRCS	0.435	0.535	0.395	0.225	0.425	0.495	0.365	0.220	0.560	0.440	0.385	0.185
		IRRCS	0.975	0.985	0.960	0.845	0.905	0.885	0.870	0.680	0.910	0.915	0.895	0.785
(100, 50)	0.75	SIS	0.935	0.915	0.855	0.415	0.875	0.905	0.795	0.385	0.890	0.910	0.850	0.850
		RRCS	0.965	0.985	0.955	0.890	0.965	0.985	0.945	0.870	0.960	0.985	0.910	0.875
		IRRCS	1	1	1	0.980	1	1	0.965	0.925	1	1	0.960	0.910
	0.5	SIS	0.935	0.905	0.810	0.390	0.795	0.845	0.740	0.355	0.855	0.890	0.730	0.380
		RRCS	0.965	0.985	0.950	0.890	0.950	0.980	0.950	0.880	0.955	0.940	0.930	0.840
		IRRCS	1	1	1	0.980	1	1	0.955	0.915	1	1	0.955	0.930
	0.25	SIS	0.815	0.880	0.680	0.305	0.680	0.740	0.585	0.260	0.760	0.860	0.720	0.370
		RRCS	0.965	0.985	0.955	0.900	0.955	0.985	0.955	0.885	0.900	0.985	0.945	0.865
		IRRCS	1	1	1	0.970	1	1	0.975	0.915	1	1	0.985	0.910

TABLE 3
(Continued)

(p, n)	λ	$\varepsilon \sim$	$N(0, 1)$				$N(0, 1)$ with 10% outliers				$t(3)$			
		Method	$\rho = 0$	0.1	0.5	0.9	0	0.1	0.5	0.9	0	0.1	0.5	0.9
(1000, 50)	0.75	SIS	0.615	0.605	0.145	0	0.515	0.490	0.130	0	0.530	0.570	0.130	0.005
		RRCS	0.750	0.705	0.485	0.230	0.640	0.650	0.435	0.215	0.710	0.640	0.435	0.180
		IRRCs	1	1	1	0.840	0.940	0.925	0.940	0.780	0.930	0.940	0.935	0.710
	0.5	SIS	0.490	0.510	0.110	0	0.366	0.370	0.080	0	0.455	0.390	0.150	0
		RRCS	0.760	0.705	0.465	0.245	0.735	0.655	0.440	0.215	0.745	0.625	0.430	0.170
		IRRCs	1	1	1	0.815	0.950	0.920	0.930	0.770	0.975	0.965	0.940	0.745
	0.25	SIS	0.200	0.215	0.035	0	0.145	0.160	0.020	0	0.155	0.210	0.055	0
		RRCS	0.755	0.695	0.470	0.240	0.675	0.665	0.440	0.215	0.755	0.615	0.375	0.215
		IRRCs	1	1	1	0.780	0.945	0.930	0.940	0.720	0.955	0.930	0.935	0.725
(1000, 70)	0.75	SIS	0.860	0.860	0.375	0.005	0.670	0.690	0.270	0.015	0.840	0.865	0.370	0.105
		RRCS	0.880	0.890	0.725	0.515	0.880	0.880	0.695	0.510	0.915	0.885	0.700	0.395
		IRRCs	1	1	1	0.970	0.960	0.945	0.935	0.910	0.970	0.985	0.930	0.915
	0.5	SIS	0.775	0.765	0.275	0.0015	0.555	0.585	0.230	0	0.760	0.750	0.280	0.0015
		RRCS	0.885	0.900	0.715	0.470	0.865	0.875	0.670	0.515	0.915	0.875	0.610	0.440
		IRRCs	1	1	1	0.950	0.955	0.945	0.935	0.900	0.955	0.950	0.915	0.875
	0.25	SIS	0.435	0.445	0.010	0	0.365	0.290	0.075	0	0.440	0.440	0.010	0
		RRCS	0.875	0.880	0.725	0.490	0.830	0.795	0.710	0.500	0.835	0.830	0.655	0.410
		IRRCs	1	1	1	0.920	0.960	0.940	0.935	0.900	0.955	0.935	0.925	0.885

TABLE 4
Proportion of SIS, gcorr, RRCS and IRRCS that include the true model for the logarithm transformation model

(p, n)	$\varepsilon \sim$	$N(0, 1)$				$N(0, 1)$ with 10% outliers				$t(3)$			
	Method	$\rho = 0$	0.1	0.5	0.9	0	0.1	0.5	0.9	0	0.1	0.5	0.9
(100, 20)	SIS	0.100	0.060	0.070	0.030	0.055	0.065	0.020	0.020	0.040	0.060	0.030	0.015
	gcorr	0.280	0.230	0.105	0.010	0.205	0.215	0.180	0.010	0.185	0.230	0.170	0.015
	RRCS	0.580	0.460	0.385	0.290	0.570	0.410	0.375	0.215	0.575	0.425	0.355	0.170
	IRRCS	1	0.975	0.975	0.715	0.875	0.870	0.875	0.560	0.905	0.875	0.840	0.580
(100, 50)	SIS	0.550	0.650	0.450	0.225	0.470	0.585	0.395	0.250	0.470	0.585	0.455	0.230
	gcorr	0.940	0.925	0.890	0.430	0.855	0.880	0.825	0.385	0.870	0.885	0.860	0.410
	RRCS	0.960	0.985	0.975	0.880	0.960	0.975	0.965	0.930	0.985	0.975	0.945	0.865
	IRRCS	1	1	1	0.980	1	1	1	0.955	0.990	1	1	0.975
(1000, 50)	SIS	0.035	0.020	0.005	0	0.015	0.005	0.020	0.010	0.020	0.010	0.005	0
	gcorr	0.420	0.415	0.285	0.015	0.385	0.405	0.025	0.005	0.340	0.410	0.265	0.010
	RRCS	0.610	0.670	0.490	0.225	0.630	0.590	0.400	0.200	0.605	0.650	0.495	0.155
	IRRCS	1	1	1	0.855	0.925	0.900	0.915	0.685	1	1	0.990	0.660
(1000, 70)	SIS	0.125	0.080	0.005	0	0.075	0.040	0.005	0	0.080	0.055	0.010	0.005
	gcorr	0.695	0.640	0.615	0.230	0.625	0.630	0.440	0.185	0.590	0.625	0.480	0.205
	RRCS	0.915	0.845	0.785	0.475	0.870	0.880	0.665	0.485	0.860	0.840	0.650	0.450
	IRRCS	1	1	1	0.940	1	1	0.960	0.930	1	1	1	0.925

The predictors are generated in the same setting as that of Fan and Song (2010), that is,

$$X_j = \frac{\varepsilon_j + a_j \varepsilon}{\sqrt{1 + a_j^2}},$$

where ε and $\{\varepsilon_j\}_{j=1}^{\lfloor p/3 \rfloor}$ are i.i.d. standard normal, $\{\varepsilon_j\}_{j=\lfloor p/3 \rfloor+1}^{\lfloor 2p/3 \rfloor}$ are i.i.d. and follow a double exponential distribution with location parameter zero and scale parameter one, and $\{\varepsilon_j\}_{j=\lfloor 2p/3 \rfloor+1}^{\lfloor p \rfloor}$ are i.i.d. and follow a mixture normal distribution with two components $N(-1, 1)$, $N(1, 0.5)$ and equal mixture proportion. The predictors are standardized to be mean zero and variance one. The constants $\{a_j\}_{j=1}^q$ are the same and chosen such that the correlation $\rho = \text{corr}(X_i, X_j) = 0, 0.2, 0.4, 0.6$ and 0.8 , among the first q predictors, and $a_j = 0$ for $j > q$. Parameter q is also related to the overall correlation in the covariance matrix.

We vary the size of the nonsparse set of coefficients as $s = 3, 6, 12, 15$ and 24 , and present the numerical results with $q = 15$ and $q = 50$. Every method is evaluated by summarizing the median minimum model size (MMMS) of the selected model and its associated RSD, which is the associated interquartile range (IQR) divided by 1.34 . The results, based on 200 replications in each scenario, are recorded in Tables 5–7. The results of SIS-based MLR, SIS-based MMLE, LASSO and SCAD in Tables 5–7 are cited from Fan and Song (2010).

From Tables 5–7, we can see that the RRCS procedure does a very reasonable job similar to the SIS proposed by Fan and Song (2010) in screening insignificant predictors, and similarly sometimes outperforms LASSO and SCAD for NP-dimensional generalized linear models.

EXAMPLE 5 (Logistic regression). This example is based on a real data set from Example 11.3 of Albright, Winston and Zappe (1999). This data set consists of 208 employees with complete information on 8 recorded variables. These variables include employee’s annual salary in thousands of dollars (Salary); educational level (EduLev), a categorical variable with categories 1 (finished school), 2 (finished some college courses), 3 (obtained a bachelor’s degree), 4 (took some graduate courses), 5 (obtained a graduate degree); job grade (JobGrade), a categorical variable indicating the current job level, the possible levels being 1–6 (6 the highest); year that an employee was hired (YrHired); year that an employee was born (YrBorn); a categorical variable with values “Female” and “Male” (Gender), 1 for female employee and 0 for male employee; number of years of work experience at another bank prior to working at the Fifth National Bank (YrsPrior); a dummy variable with value 1 if the employee’s job is computer related and value 0 otherwise (PCJob). Such a data set had been analyzed by Fan and Peng

TABLE 5
The MMMS and associated RSD (in parenthesis) of the simulated examples for logistic regressions when $p = 40,000$

ρ	n	SIS-MLR	SIS-MMLE	RRCS	n	SIS-MLR	SIS-MMLE	RRCS
Setting 1, $q = 15$								
$s = 3, \beta = (1, 1.3, 1)^T$					$s = 6, \beta = (1, 1.3, 1, \dots)^T$			
0	300	3 (1)	3 (1)	3 (0.74)	300	47 (164)	50 (170)	56 (188.05)
0.2	200	3 (0)	3 (0)	3 (0)	300	6 (0)	6 (0)	6 (0.74)
0.4	200	3 (0)	3 (0)	3 (0)	300	7 (1)	7 (1)	7 (1.49)
0.6	200	3 (1)	3 (1)	3 (0.74)	300	8 (1)	8 (2)	8 (2.23)
0.8	200	4 (1)	4 (1)	4 (2)	300	9 (3)	9 (3)	9 (2.23)
$s = 12, \beta = (1, 1.3, \dots)^T$					$s = 15, \beta = (1, 1.3, \dots)^T$			
0	500	297 (589)	302.5 (597)	298 (488)	600	350 (607)	359.5 (612)	359.5 (657.08)
0.2	300	13 (1)	13 (1)	13 (1.49)	300	15 (0)	15 (0)	15 (0)
0.4	300	14 (1)	14 (1)	14 (0.74)	300	15 (0)	15 (0)	15 (0)
0.6	300	14 (1)	14 (1)	14 (1.49)	300	15 (0)	15 (0)	15 (0)
0.8	300	14 (1)	14 (1)	14 (0.74)	300	15 (0)	15 (0)	15 (0)
Setting 2, $q = 50$								
$s = 3, \beta = (1, 1.3, 1)^T$					$s = 6, \beta = (1, 1.3, 1, \dots)^T$			
0	300	3 (1)	3 (1)	3 (0.74)	500	6 (1)	6 (1)	6 (2)
0.2	300	3 (0)	3 (0)	3 (0)	500	6 (0)	6 (0)	6 (0)
0.4	300	3 (0)	3 (0)	3 (0)	500	6 (1)	6 (1)	7 (1.49)
0.6	300	3 (1)	3 (1)	3 (1)	500	8.5 (4)	9 (5)	8 (3.73)
0.8	300	5 (4)	5 (4)	5 (3.73)	500	13.5 (8)	14 (8)	15 (7.46)
$s = 12, \beta = (1, 1.3, \dots)^T$					$s = 15, \beta = (1, 1.3, \dots)^T$			
0	600	77 (114)	78.5 (118)	95 (115)	800	46 (82)	47 (83)	46 (83.88)
0.2	500	18 (7)	18 (7)	19 (6)	500	26 (6)	26 (6)	27 (8.20)
0.4	500	25 (8)	25 (10)	26 (9.70)	500	34 (7)	33 (8)	33 (8.39)
0.6	500	32 (9)	31 (8)	32 (9)	500	39 (7)	38 (7)	38 (6.71)
0.8	500	36 (8)	35 (9)	39 (7.46)	500	40 (6)	42 (7)	42 (6.15)

TABLE 6
The MMMS and associated RSD (in parenthesis) of the simulated examples for logistic regressions when $p = 5000$ and $q = 15$

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD	RRCS
$s = 3, \beta = (1, 1.3, 1)^T$						
0	300	3 (0)	3 (0)	3 (1)	3 (1)	3 (0)
0.2	300	3 (0)	3 (0)	3 (0)	3 (0)	3 (0)
0.4	300	3 (0)	3 (0)	3 (0)	3 (0)	3 (0)
0.6	300	3 (0)	3 (0)	3 (0)	3 (1)	3 (0)
0.8	300	3 (1)	3 (1)	4 (1)	4 (1)	3 (1.49)
$s = 6, \beta = (1, 1.3, 1, 1.3, 1, 1.3)^T$						
0	300	12.5 (15)	13 (6)	7 (1)	6 (1)	12 (24.62)
0.2	300	6 (0)	6 (0)	6 (0)	6 (0)	6 (0.18)
0.4	300	6 (1)	6 (1)	6 (1)	6 (0)	7 (1.49)
0.6	300	7 (2)	7 (2)	7 (1)	6 (1)	8 (1.49)
0.8	300	9 (2)	9 (3)	27.5 (3725)	6 (0)	9 (2.23)
$s = 12, \beta = (1, 1.3, \dots)^T$						
0	300	297.5 (359)	300 (361)	72.5 (3704)	12 (0)	345 (522)
0.2	300	13 (1)	13 (1)	12 (1)	12 (0)	13 (1.49)
0.4	300	14 (1)	14 (1)	14 (1861)	13 (1865)	14 (0.74)
0.6	300	14 (1)	14 (1)	2552 (85)	12 (3721)	14 (1)
0.8	300	14 (1)	14 (1)	2556 (10)	12 (3722)	14 (0.74)
$s = 15, \beta = (3, 4, \dots)^T$						
0	300	479 (622)	482 (615)	69.5 (68)	15 (0)	629.5 (821)
0.2	300	15 (0)	15 (0)	16 (13)	15 (0)	15 (0)
0.4	300	15 (0)	15 (0)	38 (3719)	15 (3720)	15 (0)
0.6	300	15 (0)	15 (0)	2555 (87)	15 (1472)	15 (0)
0.8	300	15 (0)	15 (0)	2552 (8)	15 (1322)	15 (0)

(2004) throughout the following linear model:

$$\begin{aligned}
 \text{Salary} = & \beta_0 + \beta_1 \text{Female} + \beta_2 \text{PCJob} + \sum_{i=1}^4 \beta_{2+i} \text{Edu}_i + \sum_{i=1}^5 \beta_{6+i} \text{JobGrd}_i \\
 (5.5) \quad & + \beta_{12} \text{YrsExp} + \beta_{13} \text{Age} + \varepsilon,
 \end{aligned}$$

where the variable YrsExp is total years of working experience, computed from the variables YrHired and YrsPrior. Fan and Peng (2004) deleted the samples with age over 60 or working experience over 30 and used only 199 samples to fit model (5.5). The SCAD-penalized least squares coefficient estimator of (5.5) is

$$\begin{aligned}
 \beta_0 = & (\beta_0, \beta_1, \dots, \beta_{13})^T \\
 = & (55.835, -0.624, 4.151, 0, -1.073, -0.914, 0, -24.643,
 \end{aligned}$$

TABLE 7
The MMMS and associated RSD (in parenthesis) of the simulated examples for logistic regressions when $p = 2000$ and $q = 50$

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD	RRCS
$s = 3, \beta = (3, 4, 3)^T$						
0	200	3 (0)	3 (0)	3 (0)	3 (0)	3 (0)
0.2	200	3 (0)	3 (0)	3 (0)	3 (0)	3 (0)
0.4	200	3 (0)	3 (0)	3 (0)	3 (1)	3 (0)
0.6	200	3 (1)	3 (1)	3 (1)	3 (1)	3 (0.74)
0.8	200	5 (5)	5.5 (5)	6 (4)	6 (4)	4 (2.4)
$s = 6, \beta = (3, -3, 3, -3, 3, -3)^T$						
0	200	8 (6)	9 (7)	7 (1)	7 (1)	8 (5.97)
0.2	200	18 (38)	20 (39)	9 (4)	9 (2)	14 (28.54)
0.4	200	51 (77)	64.5 (76)	20 (10)	16.5 (6)	72 (76.60)
0.6	300	77.5 (139)	77.5 (132)	20 (13)	19 (9)	84.5 (122.94)
0.8	400	306.5 (347)	313 (336)	86 (40)	70.5 (35)	249.5 (324.62)
$s = 12, \beta = (3, 4, \dots)^T$						
0	600	13 (6)	13 (7)	12 (0)	12 (0)	13 (3.90)
0.2	600	19 (6)	19 (6)	13 (1)	13 (2)	16.5 (4)
0.4	600	32 (10)	30 (10)	18 (3)	17 (4)	23 (7)
0.6	600	38 (9)	38 (10)	22 (3)	22 (4)	29 (8.95)
0.8	600	38 (7)	39 (8)	1071 (6)	1042 (34)	35 (8)
$s = 24, \beta = (3, 4, \dots)^T$						
0	600	180 (240)	182 (238)	35 (9)	31 (10)	190.5 (240.48)
0.2	600	45 (4)	45 (4)	35 (27)	32 (24)	40 (5)
0.4	600	46 (3)	47 (2)	1099 (17)	1093 (1456)	45 (4.40)
0.6	600	48 (2)	48 (2)	1078 (5)	1065 (23)	47 (3)
0.8	600	48 (1)	48 (1)	1072 (4)	1067 (13)	47 (2.98)

$$-22.818, -18.803, -13.859, -7.770, 0.193, 0)^T.$$

For this data set, we consider a larger artificial model as a full model with additional predictors:

$$Y_j = \beta_0 + \sum_{i=1}^{13} \beta_i X_{ij} + \sum_{i=14}^{[2p/5]} \beta_i X_{ij} + \sum_{[2p/5]+1}^p \beta_i X_{ij} + \sigma \varepsilon_j, \quad j = 1, \dots, n,$$

where we set $(\beta_0, \beta_1, \dots, \beta_{13})^T = \beta_0$ that is identical to that of (5.5) above by Fan and Peng (2004), and set $\beta_i = 0$, for i with $13 < i \leq p$. Hence, X_{3j}, X_{6j}, X_{13j} and X_{ij} , $13 < i \leq p$, are insignificant covariates, whose corresponding coefficients are zero. The data are generated as follows. $(X_{1j}, \dots, X_{13j}, j = 1, \dots, n)$ are corresponding to the covariates in (5.5) and resampled from those 199 real data without replacement. For each i , $X_{ij}, 14 \leq i \leq [2p/5]$, are generated independently from the Bernoulli distribution with success

probability p_i^* where p_i^* is independently random sampled from the uniform distribution under the interval $[0.2, 0.8]$, and X_{ij} , $[2p/5] + 1 \leq i \leq p$, are generated independently from the standard normal distribution. Further, the noises ε_j , $1 \leq j \leq n$, are, respectively, generated from the normal distribution with zero mean and the standard error $\sigma = 1, 2, 3$.

To compare the performance of different methods, we set the sample size n to be 180, and, respectively, consider the different dimensions $p = 200, 400, 600$ and 1000. Consider the different sizes of $d_n = 15, 30, 60, 120$ and 179 predictors for the sure screening by the three different methods: RRCS, SIS and the generalized correlation rank method (gcorr) proposed by Hall and Miller (2009). Then we compute the proportion of the models that include the true one, which are selected by RRCS, SIS and gcorr, respectively. The experiment is repeated 200 times and the results are reported in Table 8 for various combinations of p and d_n .

From Table 8, we can see that the RRCS procedure works well in screening out insignificant predictors when there are the categorical covariates. In contrast, the SIS and gcorr methods almost cannot choose the true model. In most of the repeated experiments, we find that there are always one or two significant predictors not being selected by the SIS and gcorr methods even when $d_n = n - 1 = 179$ predictors are selected.

For SIS, such a result is consistent with the numerical study of Example 2 in Fan, Feng and Song (2011). With complex correlation structure among predictors and the response, SIS cannot work well. As for the generalized correlation screening method, its computation is complicated, especially because it has to use different methods to, respectively, calculate the generalized coefficients between the response and both categorial and continuous predictors. The variation of those coefficient estimations would be different, and make that the final sure screening results are not as stable as RRCS and SIS are.

5.2. *Application to cardiomyopathy microarray data.* Please see the supplementary material for the paper [Li et al. (2012)].

6. Concluding remarks. This paper studies the sure screening properties of robust rank correlation screening (RRCS) for ultra-high dimensional linear regression models and transformation regression models. The method is based on the Kendall τ rank correlation, which is a robust correlation measurement between two random variables and is invariant to strictly monotonic transformation. Our results discover the relationship between the Pearson correlation and the Kendall τ rank correlation under certain conditions. It suggests that the Kendall τ rank correlation can be used to replace the Pearson correlation such that the sure screening is applicable not only to linear regression models but also to more general nonlinear regression models.

In both the theoretical analysis and the numerical study, RRCS has been shown to be capable of reducing the exponentially growing dimensionality

TABLE 8
For Example 5: the proportion of RRCS, SIS and gcorr that include the true model

d_n	Method	$\sigma = 1$				$\sigma = 2$				$\sigma = 3$			
		$p = 200$	400	600	1000	200	400	600	1000	200	400	600	1000
15	RRCS	0.280	0.080	0	0	0.085	0	0	0	0.005	0	0	0
	SIS	0	0	0	0	0	0	0	0	0	0	0	0
	gcorr	0	0	0	0	0	0	0	0	0	0	0	0
30	RRCS	0.955	0.765	0.425	0.165	0.685	0.255	0.085	0.020	0.210	0.030	0.005	0
	SIS	0	0	0	0	0	0	0	0	0	0	0	0
	gcorr	0	0	0	0	0	0	0	0	0	0	0	0
60	RRCS	1	0.990	0.915	0.735	0.965	0.765	0.490	0.275	0.620	0.310	0.070	0.025
	SIS	0	0	0	0	0	0	0	0	0.005	0	0	0
	gcorr	0	0	0	0	0	0	0	0	0	0	0	0
120	RRCS	1	1	0.995	0.990	0.985	0.995	0.885	0.665	0.920	0.670	0.410	0.215
	SIS	0.045	0	0	0	0.070	0	0	0	0.125	0.005	0	0
	gcorr	0	0	0	0	0	0	0	0	0.050	0	0	0
179	RRCS	1	1	1	0.995	1	1	0.965	0.860	0.970	0.865	0.640	0.410
	SIS	0.670	0	0	0	0.660	0.010	0	0	0.715	0.015	0	0
	gcorr	1	0	0	0	1	0	0	0	1	0	0	0

of the model to a value smaller than the sample size. It is also robust against the error distribution. An iterative RRCS (IRRCs) has been also proposed to enhance the performance of RRCS for more complicated ultra-high dimensional data.

Some issues deserve further study. From Fan and Song (2010), it is easy to know that the sure screening properties of MMLE for generalized linear models really depend on $\text{Cov}(X_k, Y), i = 1, 2, \dots, n$. Hence, it is an interesting problem to determine whether the relationship between the Pearson correlation and the Kendall τ rank correlation can be identified for generalized linear models. If this can be done, the sure screening properties of RRCS for generalized linear models can also be studied theoretically. Note that the conditions required are much weaker than SIS needs. Thus, it would be of interest to determine whether robust LASSO, SCAD or other penalized methods can be defined when the idea described herein is applied.

APPENDIX: PROOFS OF THEOREMS

Please see the supplementary material for the paper [Li et al. (2012)].

Acknowledgments. Most work of this paper was finished independently by the second author or under his guidance and suggestion. The authors would like to thank Professor Jianqing Fan for his valuable suggestions and constructive discussion with the second author that improve the presentation and the results of the paper. The authors would like to thank the Editor, an Associate Editor and the referees for their helpful comments that led to an improvement of an earlier manuscript.

SUPPLEMENTARY MATERIAL

Supplement to “Robust rank correlation based screening”

(DOI: [10.1214/12-AOS1024SUPP](https://doi.org/10.1214/12-AOS1024SUPP); .pdf). Application to Cardiomyopathy microarray Data and the proofs of Theorems 1–3 and Proposition 1 require some technical and lengthy arguments that we develop in this supplement.

REFERENCES

- ALBRIGHT, S. C., WINSTON, W. L. and ZAPPE, C. J. (1999). *Data Analysis and Decision Making with Microsoft Excel*. Duxbury, Pacific Grove, CA.
- BICKEL, P. J. and DOKSUM, K. A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76** 296–311. [MR0624332](#)
- BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **26** 211–252. [MR0192611](#)
- CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *Ann. Statist.* **35** 2313–2351. [MR2382644](#)
- CARIO, M. C. and NELSON, B. L. (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Dept. Industrial Engineering and Management Sciences, Northwestern Univ., Evanston, IL.

- CARROLL, R. J. and RUPPERT, D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall, New York. [MR1014890](#)
- CHANNOUF, N. and L'ECUYER, P. (2009). Fitting a normal copula for a multivariate distribution with both discrete and continuous marginals. In *Proceedings of the 2009 Winter Simulation Conference* 352–358.
- COOK, R. D. and WEISBERG, S. (1991). Discussion with “Sliced inverse regression for dimension reduction,” by K. C. Li. *J. Amer. Statist. Assoc.* **86** 328–332. [MR1137117](#)
- DONOHU, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. In *Aide-Memoire of a Lecture at AMS Conference on Math Challenges of 21st Century*.
- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–451. [MR2060166](#)
- FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557. [MR2847969](#)
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. [MR1946581](#)
- FAN, J. and LI, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *International Congress of Mathematicians. Vol. III* (M. SANZ-SOLE, J. SORIA, J. L. VARONA and J. VERDERA, eds.) 595–622. Eur. Math. Soc., Zürich. [MR2275698](#)
- FAN, J. and LV, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. [MR2530322](#)
- FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. [MR2640659](#)
- FAN, J. and LV, J. (2011). Non-concave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. [MR2849368](#)
- FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. [MR2065194](#)
- FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional variable selection: Beyond the linear model. *J. Mach. Learn. Res.* **10** 1829–1853. [MR2550099](#)
- FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. [MR2766861](#)
- FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35** 109–148.
- GHOSH, S. and HENDERSON, S. G. (2003). Behavior of the NORTA method for correlated random vector generation as the dimension increases. *ACM Transactions on Modeling and Computer Simulation* **13** 276–294.
- HALL, P. and MILLER, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.* **18** 533–550. [MR2751640](#)
- HAN, A. K. (1987). Nonparametric analysis of a generalized regression model. The maximum rank correlation estimator. *J. Econometrics* **35** 303–316. [MR0903188](#)
- HUANG, J., HOROWITZ, J. L. and MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613. [MR2396808](#)
- HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. Wiley, Hoboken, NJ. [MR2488795](#)
- KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93.
- KENDALL, M. G. (1949). Rank and product-moment correlation. *Biometrika* **36** 177–193. [MR0034990](#)
- KENDALL, M. G. (1962). *Rank Correlation Methods*, 3rd ed. Griffin & Co, London.

- KLAASSEN, C. A. J. and WELLNER, J. A. (1997). Efficient estimation in the bivariate normal copula model: Normal margins are least favourable. *Bernoulli* **3** 55–77. [MR1466545](#)
- LI, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–342. [MR1137117](#)
- LI, G., PENG, H. and ZHU, L. (2011). Nonconcave penalized M -estimation with a diverging number of parameters. *Statist. Sinica* **21** 391–419. [MR2796868](#)
- LI, G. R., PENG, H., ZHANG, J. and ZHU, L. X. (2012). Supplement to “Robust rank correlation based screening.” DOI:[10.1214/12-AOS1024SUPP](#).
- LIN, H. and PENG, H. (2013). Smoothed rank correlation of the linear transformation regression model. *Comput. Statist. Data Anal.* **57** 615–630.
- LV, J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. [MR2549567](#)
- NELSEN, R. B. (2006). *An Introduction to Copulas*, 2nd ed. Springer, New York. [MR2197664](#)
- PITT, M., CHAN, D. and KOHN, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93** 537–554. [MR2261441](#)
- SEN, P. K. (1968). Estimates of the regression coefficient based on Kendall’s tau. *J. Amer. Statist. Assoc.* **63** 1379–1389. [MR0258201](#)
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. [MR1379242](#)
- VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. [MR2396809](#)
- WACKERLY, D. D., MENDENHALL, W. and SCHEAFFER, R. L. (2002). *Mathematical Statistics with Applications*. Duxbury, Pacific Grove, CA.
- WANG, H. (2012). Factor profiled sure independence screening. *Biometrika* **99** 15–28. [MR2899660](#)
- XU, P. R. and ZHU, L. X. (2010). Sure independence screening for marginal longitudinal generalized linear models. Unpublished manuscript.
- ZHU, L. P., LI, L. X., LI, R. Z. and ZHU, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1474. [MR2896849](#)
- ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. [MR2279469](#)
- ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. [MR2137327](#)
- ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36** 1509–1566. [MR2435443](#)

G. LI
COLLEGE OF APPLIED SCIENCES
BEIJING UNIVERSITY OF TECHNOLOGY
BEIJING 100124
CHINA
E-MAIL: ligaorong@gmail.com

H. PENG
L. ZHU
DEPARTMENT OF MATHEMATICS
HONG KONG BAPTIST UNIVERSITY
HONG KONG, CHINA
E-MAIL: hpeng@math.hkbu.edu.hk
lzhu@math.hkbu.edu.hk

J. ZHANG
SHEN ZHEN-HONG KONG JOINT RESEARCH CENTER
FOR APPLIED STATISTICS
SHENZHEN UNIVERSITY
SHENZHEN 518060
CHINA
E-MAIL: zhangjunstat@gmail.com